Potential for Counterfactual Explanations to Support Digitalized Plant Operations

Abhijit Bhakte,a Rajagopalan Srinivasan,a,b,\*

aDepartment of Chemical Engineering, Indian Institute of Technology Madras, Chennai 600036, India

bAmerican Express Lab for Data Analytics, Risk & Technology, Indian Institute of Technology Madras, Chennai 600036, India

\*[raj@iitm.ac.in](file:///C:\Users\Admin\Desktop\ESCAPE\raj@iitm.ac.in)

Abstract

The advent of digitalization in the chemical industry has begun a new era of technological advancements, where artificial intelligence (AI) methods play a pivotal role in diverse applications, from process optimization to product development. However, a major challenge persists – operators and stakeholders often find it challenging to comprehend the outputs provided by AI models due to their inherent complexity. Counterfactual explanations (CFs) are emerging as a promising approach to demystify AI predictions. A counterfactual explanation tells you what changes could have been made in a situation to get a diverse outcome. CF is rooted in causal inference and thus has the potential to provide insights, especially related to operations in complex chemical processes. Using process monitoring as an example application, this paper demonstrates that CF can be used for various use cases – aiding model developers during the training phase of AI applications, empowering plant operators to better interpret AI predictions, and training new personnel effectively.

**Keywords**: Explainable Artificial Intelligence, Deep Learning, Chemical Industry, Process Operations

* 1. Introduction

The advent of Industry 4.0, improved instrumentation, and ubiquitous sensor data play a pivotal role in the development of AI. However, despite the better performance, these algorithms often lack interpretability in their decision-making processes. Therefore, domain experts in the chemical industry may find it hard to understand or trust the outputs and recommendations provided by AI-based models. This poses a significant barrier to the adoption of AI in real-world plant operations, where safety and reliability are paramount. In response to the broad interpretability challenge of AI models, there is a recent emergence of Explainable Artificial Intelligence (XAI).

XAI offers tools and techniques to generate high-quality, interpretable, intuitive, human-understandable explanations for AI prediction. The common ways to enable humans to interpret any situation are visual explanation, explanation-by-simplification, rule-based explanation, feature relevance explanation, and explanation-by-examples (Arrieta et al., 2020). In the chemical industry, where sensor data is predominant, the major utilization of feature relevance and explanation-by-simplification is observed. For example, Bhakte et al. (2022) proposed an attribution-based XAI method that identifies key variables responsible for fault occurrence as a feature relevance method. Bhakte et al. (2023) proposed the Limit-based Explanations for Monitoring (LEMON) method, emphasizing model simplification with the process knowledge. Explanation-by-example is a method that involves the extraction of data examples closely tied to the results generated by a specific model. These techniques include: 1) Prototype explanations that involves “instances other than predefined inputs that effectively represent the model predictions.” 2) Counterfactual explanations involve the "smallest change to the feature value that changes the prediction to a predefined output" (Molnar, 2022).

Counterfactual explanations, rooted in casual inference, enable stakeholders to comprehend why an AI model generated a particular output, what factors contributed to it, and how changes to these factors can affect the results. CF has been applied in various domains, from finance to healthcare, yet its potential remains underexplored in chemical engineering. Specifically, within process operations, the application of CFs enables us to envision what would occur in various scenarios. While factual explanations aid in logical reasoning for a process operator, counterfactuals offer additional insights beyond the known facts. Hence, they are more informative and facilitate creative problem-solving. Thus, CFs can be used for numerous applications, such as predicting the remaining useful life and dynamic planning of operations. This work describes a methodology to generate CF explanations and explores its potential to help decision-making during process monitoring.

* 1. Methodology

Consider a chemical process that is being monitored by a DNN model . The DNN is trained on a dataset of historic samples with labels . Here, is the training samples with measurement variables. Given an online process sample , the model outputs the probability of process states, i.e., ]. The end-user is provided with the most probable state . To enable the operator to understand the output we propose a two-step methodology that generates the explanation-by-example.

***Step 1: Variable Selection***

The trained DNN is deployed for online process monitoring. Now, understanding the AI prediction in time, given process variables, becomes challenging. Hence, if an operator has given the subset of variables responsible for the model output, this makes his job easier and aids in fast process recovery in abnormal situations. To address this, Integrated Gradient methodology in (Bhakte et al., 2022) is used for feature selection. The IG is a gradient-based attribution motivated by the concept of Shapley value, a technique to explain DNNs predictions by attributing them to the neural network's inputs . Now, to select the top highly attributed variables, we use the concept of inseparability defined as the “ratio of the attribution of the variable to that of the next highest attributed variable.” Here, represents variable attribution wheras index represents the ranking of variables' attribution value sorted in descending order; a low value represents that both variables at position and are important, whereas represents only the variable at position is important. Therefore, a threshold value must be chosen to display the key variables to the operator. Here, we have only key variables selected, but more insights can be obtained if distinct ‘if-else’ scenarios are provided to the operator.

***Step 2: Counterfactual Generation***

The process state and the subsequent classes with the highest probabilities obtained from DNN output are stored. Next, the nearest neighbors are searched from historic data that belongs to class . This nearest neighbor is called prototype and calculated using the loss function .

|  |  |
| --- | --- |
|  | (1) |

The training dataset is utilized to select the nearest neighbors as it automatically incorporates complex causal relationships amongst the process variables. Here, prototypes help the operator to see how the real-time sample best fits the predicted state . In contrast, CFs for the subsequent classes with the higher probabilities after are obtained using the loss function .

|  |  |
| --- | --- |
|  | (2) |

These CFs help the operator understand what distinguishes the real-time sample from the other classes. Next, the prototypes and CFs for the key variables selected during Step 1 are presented to the operator. This helps the operator to understand the DNN output by focusing on key variables responsible for process state prediction. The proposed methodology is demonstrated through a CSTR case study.

* 1. Case Study: Continuous Stirred Tank Reactor

This section illustrates the proposed methodology using a jacketed CSTR process featuring the irreversible liquid-phase decomposition of reactant A into B, driven by first-order kinetics. The reaction is exothermic; therefore, a cooling jacket is employed to maintain the reactor temperature at . The process consists of five measurement variables: reactor inflow , reactor outflow , outlet concentration , reactor temperature , Reactor height , and Coolant temperature . The process consists of two steady states, NS-1 and NS-2. It is susceptible to five faults, i.e., Catalyst deactivation (F-1), Heat exchanger fouling (F-2), Ramp decrease in (F-3), Ramp increase in (F-4), and Step increase in the inlet (F-5). A DNN model is trained for monitoring using the aforementioned six variables. The training dataset consists of 120 samples from six simulation runs where faults were introduced at sample . In the testing dataset, each fault test simulation run consisted of 100 samples where faults were introduced at sample . In this work, we seek to provide an explanation-by-example for the outputs generated by DNN. To achieve this, a DNN model (6–50–25–12–7) is trained using the training dataset, and the resultant model is subsequently deployed for real-time monitoring. The proposed methodology is then employed to explain the DNN’s predictions.

In this study, we consider test run 1 as an illustrative example. During test run 1, a fault occurred at sample causing a decrease in coolant temperature . This increases the heat removal rate, resulting in a decrease in reactor temperature . The decrease in the reactor temperature causes a drop in the reaction rate, leading to an increase in the concentration of unreacted A (. The DNN correctly flags F-3 (Ramp decrease in ) with a diagnosis delay of 11 samples. To interpret the DNN predictions, Prototypes and CFs are generated through the proposed methodology. Figure 1 provides a visual representation of this analyses, specifically focusing on key variables, i.e., () identified through variable selection. Each plot in Figure 1 represents a real-time sample, 10 preceding samples (depicted as lines), 10 prototype samples (depicted as triangles), and 10 CF samples from two subsequent classes with the highest prediction probabilities (represented by squares). Our analysis yields several insights, offering valuable information for operators to interpret and act upon the DNN predictions:

***Steady-state insights****:* These insights involve understanding DNN when the process operates within the normal operating range. During test run 1, all process variables remained within normal operating range until , and the DNN model classifies these samples correctly. To understand the DNN predictions, Prototypes and CF generated through the proposed methodology are analyzed at sample . At , the model predicted the given sample as normal with a probability of 0.97; it also flags the class F-1 with 0.01 probability as the next likely candidate. These probabilities clearly indicate that the current state as normal. This is evident in Figure 1 (a) & (b), where the present samples align within the normal class cluster (prototypes). In contrast, clusters corresponding to abnormal samples, i.e., F-1 and F-4 (CF w.r.t. current samples) are far away.

|  |  |
| --- | --- |
| (a) | (b) |

Figure 1 Counterfactuals and prototypes for test run 1 at t=25

***Target-based insights:*** These refer to actionable information provided to operators for returning the system to a desired normal state. In the case study, the emergence of a fault initiates a gradual escalation at , resulting in the misclassification of the initial five samples as normal by the DNN. Subsequently, at , the DNN identifies the fault as F-1. Analysis of the sample at , the DNN predicted the given sample as F-1 with a probability of 0.42; it also flags the class F-3 and NS-1 with 0.34 and 0.24 probability as the next likely candidate. The corresponding prototypes and CFs plotted in Figure 2 (a) & (b) demonstrate the sample's proximity to the decision boundaries of F-3 and NS-1, representing diagnostic uncertainty. In such a situation, target-based insights suggest that to restore normal steady state-1 (NS-1), and need to be increased to 345.20°C and 402.10°C, respectively. These insights provide actionable guidance for operators to rectify deviations from the desired state.

***Pattern-based insights:*** These involve the analysis of observed patterns in data, particularly focusing on the historical behavior of variables within a system. In the case study, examining the 10 prototypes of fault class F-1 and counterfactuals of class F-3 at (see Figure 2), if alone changes with time, it confirms the F-1 state. In contrast, if a 1° change in changes 0.0015 mol/L of it confirms the F-3 state. This knowledge is not explicitly given to the DNN but extracted from the training data in the vicinity of the real-time sample. Utilizing these counterfactuals and prototype explanations proves instrumental for operators navigating diagnostic uncertainties. Utilizing these explanations enhances operators' ability to interpret evolving states more comprehensively.

|  |  |
| --- | --- |
| (a) | (b) |

Figure 2. Counterfactuals and prototypes for test run 1 at t=38

***Counter Insights:*** It explains instances where a model deviates from an expected outcome. In the case study, the fault state F-3 is detected by DNN at sample . One such sample at predicted as F-3 by DNN with a probability of 0.78; it also flags the class NS-2 with 0.2 probability as the next likely candidate. The CF and prototypes are generated to explain the DNN prediction at as shown in Figures 3 (a) & (b). At , why the current sample is not predicted as F-1, even though is within the fault range? The counter insight reveals that the significant drop in and indicates a different fault F-3. Understanding these counter insights is essential for operators navigating diagnostic uncertainties and refining their understanding of the fault detection process.

***Nearest Optimal Insights*:** These insights guide the operator towards the most feasible steady state when returning the desired normal steady state is unattainable. In the case study, if the operator aims to return to normal steady state-1 (NS-1) at (see Figure 3), he can follow the distinct control moves to reach that state. However, if returning to NS-1 is impossible, the operator is provided with the nearest normal that can be operated, i.e., normal steady state-2 (NS-2) in the current scenario. Operating the system under NS-2 ensures stability and minimizes potential risks.

|  |  |
| --- | --- |
| (a) | (b) |

Figure 3.Counterfactuals and prototypes for test run 1 at t=85

In summary, The case study showcases five key potential insights obtained from explanation-by-examples: *steady-state insights* for normal state, *target-based insights* for returning to normal, *pattern-based insights* for understanding evolving states, *counter insights* explaining why samples are not predicted as specific faults, and *nearest optimal insights* guiding operators to an operable normal state. These insights help the operator to understand AI decisions more intuitively and aid the process recovery.

* 1. Conclusions

This work introduces the potential of explanations-by-example in the chemical industry, addressing the interpretability challenges associated with AI model results. The proposed methodology combines variable selection and counterfactual generation to enhance the understanding of AI predictions for the operator. Using counterfactuals in the CSTR Case study, we presented the distinct potential of prototype and counterfactual to provide insights, i.e., steady-state, target-based, pattern-based, counter, and nearest optimal insights. This aids in decision-making, process recovery, and effective training of plant personnel. This innovative use of CFs holds significant potential for developing trust and facilitating the adoption of AI in real-world chemical industries. In the future, we will focus on providing insights as well as systematic mitigation strategies to the operator for process recovery.

References

Arrieta, A.B., Díaz-Rodríguez, N., Ser, J.D., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bhakte, A., Chakane, M., Srinivasan, R., 2023. Alarm-based Explanations of Process Monitoring Results from Deep Neural Networks. Computers & Chemical Engineering 179, 108442. https://doi.org/10.1016/j.compchemeng.2023.108442

Bhakte, A., Pakkiriswamy, V., Srinivasan, R., 2022. An explainable artificial intelligence-based approach for interpretation of fault classification results from deep neural networks. Chemical Engineering Science 250, 117373. https://doi.org/10.1016/j.ces.2021.117373

Molnar, C., 2022. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.